Integration of information and volume visualization for analysis of cell lineage and gene expression during embryogenesis

Andrej Cedilnik, Jeffrey Baumes, Luis Ibanez, Sean Megason, Brian Wylie

Andrej Cedilnik, Jeffrey Baumes, Luis Ibanez, Sean Megason, Brian Wylie, "Integration of information and volume visualization for analysis of cell lineage and gene expression during embryogenesis," Proc. SPIE 6809, Visualization and Data Analysis 2008, 68090J (28 January 2008); doi: 10.1117/12.768014

SPIE.

Event: Electronic Imaging, 2008, San Jose, California, United States

Integration of Information and Volume Visualization for Analysis of Cell Lineage and Gene Expression during Embryogenesis

Andrej Cedilnik[a], Jeffrey Baumes[b], Luis Ibanez[b], Sean Megason[c], and Brian Wylie[d]

[a]TiVo, Inc., Alviso, CA, USA;
[b]Kitware, Inc., Clifton Park, NY, USA;
[c]California Institute of Technology, Pasadena, CA, USA;
[d]Sandia National Laboratories, Albuquerque, NM, USA

ABSTRACT

Dramatic technological advances in the field of genomics have made it possible to sequence the complete genomes of many different organisms. With this overwhelming amount of data at hand, biologists are now confronted with the challenge of understanding the function of the many different elements of the genome. One of the best places to start gaining insight on the mechanisms by which the genome controls an organism is the study of embryogenesis.

There are multiple and inter-related layers of information that must be established in order to understand how the genome controls the formation of an organism. One is cell lineage which describes how patterns of cell division give rise to different parts of an organism. Another is gene expression which describes when and where different genes are turned on. Both of these data types can now be acquired using fluorescent laser-scanning (confocal or 2-photon) microscopy of embryos tagged with fluorescent proteins to generate 3D movies of developing embryos. However, analyzing the wealth of resulting images requires tools capable of interactively visualizing several different types of information as well as being scalable to terabytes of data.

This paper describes how the combination of existing large data volume visualization and the new Titan information visualization framework of the Visualization Toolkit (VTK) can be applied to the problem of studying the cell lineage of an organism. In particular, by linking the visualization of spatial and temporal gene expression data with novel ways of visualizing cell lineage data, users can study how the genome regulates different aspects of embryonic development.

Keywords: Information Visualization, Volume Visualization, Large Data Visualization, Cell Lineage

1. INTRODUCTION

This paper describes an application for analyzing cell lineage and gene expression during embryogenesis. The approach taken in this application is to combine spatial visualization, such as volume rendering of confocal microscopy images from *Caenorhabditis elegans* (*C. elegans*) embryos together with information visualization displays, such as tree visualization of the lineage.

Section 2 is a quick summary of related work in the areas of Cell Lineage, Information Visualization, and Information and Scientific Visualization. Section 3 describes the problem that the application is trying to solve. It also defines the concepts of Cell Lineage and relationship between the Cell Lineage and visualization. Section 4 describes the application and addresses the reproducibility of the results presented in this paper. Finally Section 5 provides information about the future development of the application and the areas where future research can be performed.

Further author information: (Send correspondence to)
Andrej Cedilnik: E-mail: andy@cedilnik.com;
Jeffrey Baumes: E-mail: jeff.baumes@kitware.com;
Luis Ibanez: E-mail: luis.ibanez@kitware.com;
Sean Megason: E-mail: megason@caltech.edu;
Brian Wylie: E-mail: bnwylie@sandia.gov

Visualization and Data Analysis 2008, edited by Katy Börner, Matti T. Gröhn, Jinah Park, Jonathan C. Roberts, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6809 68090J, © 2008 SPIE-IS&T · 0277-786X/08/$18

2. RELATED WORK

Visualization for biological applications has evolved in many different forms. Currently biologists communicate by using a set of *de facto* standards that have flourished in their technical literature. One of the commonly used information visualization techniques relates to the display of genealogical trees. This is important in the fields of ecology, evolution and cell biology, in particular for the study of cell lineages in the latter.

Combining genealogical trees information visualization with 3D visualization of confocal microscopy involves the use of a collective set of technologies related to computer graphics, image analysis and visualization. The scalability of these technologies is put to the test with the large amounts of data that biological research can produce.

The following sections describe related work in the areas of cell lineage determination, information visualization and large data processing.

2.1 Cell Lineage

Cell lineage refers to the pattern of cell divisions during embryonic development. It is an important and long studied problem in biology. This is especially true for *C. elegans* a microscopic roundworm that became a popular model organism for biological research in part due to its invariant cell lineage. Life begins in *C. elegans* with the fertilized egg cell termed P0. The fertilized egg undergoes a a number of rounds of cell division, cell death, and cell differentiation leading to the formation of the adult organism. Importantly, in *C. elegans* this pattern of division, death, and differentiation is the same in all organisms which allowed the complete lineage from the 1 cell fertilized egg to the 959 cell adult organism to be established through repeated observation of different individuals and manual reconstruction.[1,2] More recently, confocal microscopy has advanced to the point that 3D movies of much of embryonic development in *C. elegans* can be captured with high enough resolution to segment and track cell lineages automatically from a single individual.[3] Software tools have also been developed to automatically annotate the segmented cells with their individual identities.[4]

In toto imaging in which high performance laser-scanning (confocal or 2-photon) microscopy is used to capture 3D movies of every cell in a developing tissue means it is now possible to extend cell lineage analysis to organisms that do not have an invariant lineage, such as vertebrates which are more medically relevant.[5] For *in toto* imaging, embryos are labeled to allow all the cells to be segmented such as through the use of a nuclear localized green fluorescent protein and a membrane localized red fluorescent protein. 3D movies are then taken on a laser-scanning microscope and special software such as GoFigure is used to segment and track the cells.[5] Importantly, transgenic organisms can be used to mark where different genes are expressed by using additional colors of fluorescent proteins.[6] Thus, both cell lineage information and gene expression information can be collected in parallel from a single organism. A significant challenge for the future is visualizing these different data types and dealing with the large image sets which can contain over 100,000 images per experiment.

The image analysis techniques that are required for identifying individual cells in confocal microscopy 3D datasets involve a combination of of noise reduction techniques, image segmentation, and pattern recognition methodologies. Most of them can be constructed based on existing functionalities available in open source toolkits such as the Insight Toolkit ITK.[7] Scaling these capabilities to large size datasets, in the orders of gigabytes, requires however to integrate into ITK the parallel commutation capabilities of the Visualization Toolkit VTK.

2.2 Information Visualization and Large Data

The development of *C. elegans* has been studied intensely by biologists, which has included visualizing it using a variety of information visualization techniques.[8] The most fundamental of these is a tree layout, where cell divisions are naturally represented by a binary tree. Trees may be displayed in a standard top-down format, or in a radial layout. An excellent example of a tree viewing application is the Interactive Tree of Life, which allows users to browse a phylogenic in standard or radial layout.[9] The Tree of Life also links in various other types of data, including images of each organism and data about genome size and when the genome was published.

In the case of embryonic development, we have volumetric data that, when combined with the tree, makes the application much more powerful. A slice through the tree at a certain point in time should be tied to the a volumetric view of the cells at that stage of development. Also, cells that are selected in the tree view should be made visible in the volume view, and vice versa.

There are a number of methodologies for linking views together. North describes an architecture for linking views by snapping them together based on scrolling and selection.[10] An alternative mechanism is proposed by Pattison and Phillips which follows the model-view-controller software pattern.[11]

There are existing visualization toolkits such as Prefuse,[12] GraphViz,[13] the InfoVis Toolkit,[14] which have capabilities specific to information visualization, but lack scientific visualization tools. VTK, whose development and use has largely been on scientific visualization, contains a Titan toolkit which provides functionality for constructing and visualizing tables, graphs and trees.[15] Titan is designed, developed, and supported by Sandia National Laboratories. Designing a collaborative software tool that works with VTK and Titan utilizes both scientific and information visualization technologies.

There are several challenges in developing the visualization application. Data processing, imaging, graph and tree layout, volume visualization, and computer interaction are required. To address this, there are several toolkits available that facilitate some or all of the features required. Example of this are the Visualization Toolkit (VTK) developed by Schroeder et al[16,17] and the Insight Toolkit (ITK). VTK provides high quality data processing, imaging, and data visualization using a simple easy to understand application programmable interface (API). In addition to its established spatial visualization capabilities, it is now capable of doing fairly elaborate information visualization using its Titan informatics toolkit. The main benefit of using VTK as a base is that the boundary between information visualization and traditional visualization can be blurred. This means that both can be used within the same application and even within the same visualization pipeline.

Even when analysing embryogenesis of simple organism, the amount of data can be extremely large. As an example, a confocal microscopy image of a zebra fish embryo results in almost a quarter billion voxel data. Visualizing this on a single computer can be extremely slow and far from interactive. In addition to this, the cell lineage this organism can be in the order of 100,000s to millions of cells. If we add the gene expression information for all 20,000 genes for each cell, just the simple processing of data can be prohibitively expensive. To solve this, the rendering and data processing of ParaView can be used. Several works,[18–20] describe the ParaView large data visualization framework. This framework allows efficient visualization of large data, by partitioning and distributing data across a cluster of computers and perform data processing as well as rendering on the local processors. The final results are composited into the image that is displayed to the user. Currently ParaView is mainly used for performing scientific visualization. There are almost no information visualization capabilities. However, since ParaView is build on top of the Visualization Toolkit, these features will eventually be available.

3. DEFINITION OF THE PROBLEM

Dramatic technological advances in the field of genomics have made it possible to sequence in short times and at relatively low cost the genomes of many different organisms. With this overwhelming amount of information at hand, biologist are now confronted with the challenge of understanding the function of the many different elements of the genome. One of the best places to start gaining insight on the mechanisms by which the genome controls an organism is the study of embryogenesis.

No matter how complex an organism is, it always starts with a single egg cell. The process by which this original cell divides and its descendents specialize in order to produce a fully functional organism is the object of study of developmental biology. Advances in molecular biology and more recently genomics have transformed this field from being based solely on descriptive observations of embryo morphology to a realm in which the mechanisms underlying development can be understood as the output of networks of interactions of genes and proteins. This exciting new approach is called systems biology.

The first layer of information that must be established when studying the development of an embryo is how different descendants of the egg cell divide and specialize in order to give origin to specific anatomical structures. Understanding the interplay of genes during embryogenesis requires to study when and where every relevant gene is expressed in the embryo during the different stages of development. Biologist can do this today by acquiring fluorescence confocal microscopy images of embryos for which specific genes have been tagged with fluorescent proteins. Analyzing the resulting information requires information tools capable of scaling to the level of Tera bytes of data and beyond.

The challenge of elucidating the cell lineage of an organism is usually attacked in multiple stages. First, the microscopy images are processed for segmenting the individual cells. Second, the cells are associated across different time points in

order to establish which cells result from the division of cells in a previous time point. Third, visualization tools must be used for analyzing the cell lineage data along with the spatial information.

This procedure has been applied to simple organisms such as the *Caenorhabditis elegans*, which produces only 1090 cells during development (the adult contains only 959 cells because 131 cells undergo programmed death during development). However, when addressing more complex organisms such as the Zebrafish the data that needs to be processed gets scaled by two or three orders of magnitude.

This paper describes the type of visualization functionalities that are needed by the community of biologist studying embryogenesis, and how many of them are being implemented in the Visualization Toolkit (VTK). Scalability is one essential requirement for software tools to satisfy the needs of data processing for larger organisms. VTK's existing parallel infrastructure has already proved to be scalable to terabytes of data and therefore provides a platform for seamlessly developing information visualization applications suitable for the study of embryogenesis.

3.1 Following Cell Lineage

The process of embryogenesis starts with the zygote cell dividing into two cells of approximately half the size of the zygote. These two cells divide each one into two daughter cells, which continue to divide. The division process is not totally symmetric. At every stage, the daughter cells have slightly different concentrations of proteins, and slightly different geometric relationships with their neighbors. These subtle composition differences, along with further exchanges of information via molecular signals between the cells, drive the process of differentiation by which the cells become specialized into particular tissues and organs. The daughter cells also arrange themselves in particular spatial patterns. For example, the first two divisions of the *C. elegans* zygote define the basic body-plan of the organism; each cell being associated to one of the anterior, posterior, ventral and dorsal axes of the worm.

The history of cellular divisions that sequentially tracks the parents of a particular cell up to its original ancestor, the zygote cell, is called *cell lineage*. Simple organisms such as the *C. elegans* worm tend to have a deterministic cell lineage, meaning that for any given embryo, the cells divide following exactly the same pattern. More complex organisms tend to have a deterministic lineage only in the early stages of embryogenesis and then, as large numbers of cells of similar type are required, they tend to generate these cells by following non-deterministic, although likely sterotypic, patterns.

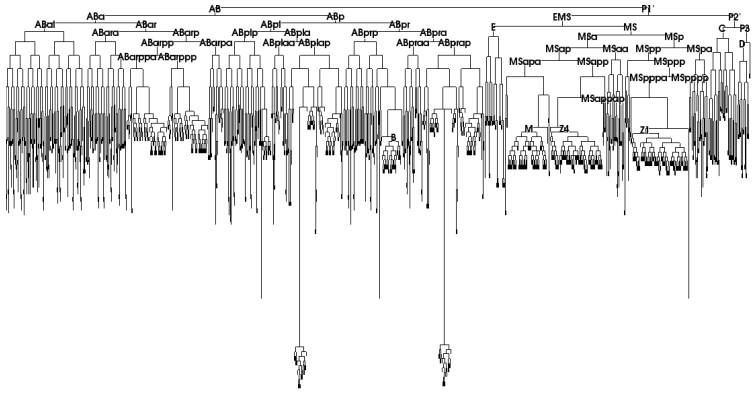

Figure 1. Complete cell lineage of *C. elegans*.

Studying the cell lineage of an organism is important because it helps in the understanding of how a particular cell defines its fate in the organism. It also helps to understand how mutations and environmental factors may perturb the embryogenesis process and result in malformed organisms. In the context of studying gene functionality, the cell lineage also provides a baseline layer upon which the behavior of genes can be laid out across space and time as the organism develops. For example, if a biologist is testing the hypothesis that a particular gene is required for the formation of the eyes, tracking the expression of this gene across the cell lineage of the embryo will permit to verify if different levels of expression of this gene may make a difference in how the eyes get formed.

Cell lineages are also extremely important in tissue-engineering applications where undifferentiated stem cells are cultivated in order to produce tissue suitable for use as grafts for particular organs. This is an important treatment for

patients suffering from severe burns and muscular degeneration, for example. In this context, understanding the lineage of the cells that are to be produced helps to select the appropriate ancestor cells and culture conditions that could be used as starting point for cultivating the tissue.

Studying the cell lineage of an organism, however, is a very challenging task and to date has only been done for a few simple organisms. It typically requires a scientist to carefully observe the development of the embryo under a microscope and to track each one of its cellular divisions. The task grows more and more challenging as the development of the embryo progresses and more and more cells occupy the field of view. The recent combined use of fluorescence confocal microscopes and automated imaging technologies has provided better tools for confronting the challenge of tracking thousands of cells over time. A recent and very promissing approach to the study of embryogenesis is to acquire 4D datasets by grabbing 3D confocal microscopy datasets repeatedly over time as the embryo develops. These datasets are then processed using image segmentation algorithms in order to identify the location of the cells at every time.[3,5,8]

Once the cells have been segmented from the images, the sequence of cellular divisions that led to the creation of each one of the cells has to be traced over time in order to reconstruct the links between each daughter cell and its parent. Current methodologies for tracing cell lineages involve a combination of automated tracing followed by supervised post-processing. The automatic tracing is based mostly on interpreting the spatial location of the cells with respect to the body-plan axes, and identifying the spatio-temporal continuity between the parent cells and its daughter cells. The supervised post-processing stage is performed by an expert interactively editing a graphical representation of the lineage structure.[3,4]

3.2 Information and Spatial Visualization

While many software systems provide tools in either spatial-temporal or information visualization, they have not been integrated in a meaningful way. Here we describe the differences between spatial and information visualization, and how they should be used together to make an effective system.

In spatial visualization, the data to be visualized has some inherent spatial information assigned to each data point (i.e., a space reference frame that users easily relate to). The challenge of spatial visualization is to transform the data in such a way as to effectively show important features in the data.

Spatial visualization algorithms may operate on scalar data (e.g. color mapping, contouring), vector data (e.g. oriented glyphs, warping, streamlines), tensor data (e.g. tensor ellipsoid glyphs), or may perform other advanced geometric or topological operations (e.g. cut planes).

With information visualization, the problem is abstracted one level further. Most of the time, there is no inherent spatial-temporal information given to information (no intuitive perceptual framework). Instead, the raw information consists of an arbitrary number of fields assigned to a set of entities, and a perceptual framework must be constructed. The data for information visualization may come from a wide variety of sources, such as financial or marketing data. In certain cases there may also be a predefined set of linkages between objects, for example in communication graphs or corporate hierarchies. These linkages may be explicit, or could be constructed implicitly (e.g. via a similarity metric between objects). As with spatial data, the information must then be filtered and transformed in order to extract the most useful information from the data.

The additional task of information visualization involves assigning a physical location to each entity. There are a wide variety of ways to perform this operation, including projections, graph layout, tree maps, and parallel coordinates. The resulting data containing physical locations may then be filtered in a variety of ways using spatial visualization filters.

One case where spatial and information visualization may be used together is when visualizing metadata. For simple experiments with a small number of datasets, the user may be able to keep track of the meaning of datasets by their filename or location on the file system. However, as the number of datasets becomes large as in a scientific experiment, the user must then rely on metadata in order to (1) find relevant dataset(s) and (2) view information about loaded datasets. This metadata, taken collectively from a number of datasets, may be considered a data source itself. If the metadata from numerous datasets are stored in a single location, such as a database, along with a pointer to the full datasets (e.g. via a filename), the system should be able to perform some additional advanced operations. First, the metadata could be read in and visualized using information visualization techniques. For example, the metadata for a series of experiments could be plotted using a parallel coordinates plot in order to view relationships between experiment runs along with interesting outliers. This metadata should be linked to the full datasets so that the user may open the relevant datasets. These datasets may then be visualized with spatial visualization algorithms.

4. METHODOLOGY

The application was implemented using VTK, Titan, and Qt. Each toolkit aided towards the ease of implementation. VTK and Titan provides several algorithms for spatial and non-spatial data processing, as well as visualizing. The application utilizes the flexibility of the toolkits to perform tasks that are usually hard in the traditional Information Visualization toolkits. For example, to define the time front, the iso-contour line was applied to the tree visualization.

All the toolkits used in this application are cross platform toolkits. As such they not only cross platform boundaries, but also domains-of-use boundaries. By leveraging the large user communities of these two toolkits, the application benefits from out-of-the-box ideas borrowed from the related fields.

This section describes the implementation and usage of the application, as well as the scalability of technology and reproducibility of this project.

4.1 Application

The basic pipeline of the application can be seen on the Figure 2. There are two distinct branches of the visualization pipeline. On one hand, the actual cell lineage is visualized using Titan module. The confocal microscopy volumetric images, on the other hand, are visualized using the more traditional volume rendering capability of VTK. The selection of the transfer function used for volume rendering as well as the selection of the time frame of the 3D data to be rendered, out of the 4D collection of confocal microscopy images, is determined by the selection of nodes within the InfoVis part of the pipeline. In this way, it is possible to navigate the cell lineage tree and to obtain 3D visualizations of the embryo at a stage that is consistent with the selections made on the cell lineage tree display.

4.1.1 Information Visualization Pipeline

The information visualization components of the application are provided by the Titan Informatics Toolkit.[15] Titan is designed, developed, and supported by Sandia National Laboratories.

Titan provides the ability to extract meaningful knowledge from large information sources through various visualization and interaction techniques. Titan has been added onto the existing capabilities of the VTK, which allows rich integration between information visualization and scientific visualization.

There are a number of data object types which may store various geometric objects such as image data, volume data, or polygonal data. Titan defines additional types for tables, trees, and graphs. For the cell lineage viewer, we use the tree structure. A tree holds a hierarchy of entities. In the case of a cell lineage, the cells are the entities and the hierarchy is naturally defined by the parent and child cells, which in effect creates a binary tree.

The tree is read in from a file and loaded into two separate views. One view is a tree widget which shows the cell hierarchy as an expandable list of cells. The other view is a tree layout view where the tree is laid out using either a standard view or a radial view. The tree is then converted to polygonal data and rendered in the same way as other data types. Using Titan we are also able to link the selection in these views.

4.1.2 Volume Visualization Pipeline

To visualize the confocal microscopy images of the embryo, volume visualization is used. The toolkits provide necessary data structure and algorithms to perform the volume visualization. The basic data structure is vtkImageData, which is a three dimensional block of data (voxels) with origin, spacing, and dimensions associated with it. There are several volume mappers. The most commonly used are fixed-point software ray-casting and GPU-based ray-casting. Depending on the hardware on which the application is running, the appropriate mapper is chosen. Figure 3 (a) shows a volume display from a confocal microscopy image of the embryo development.

In order to dynamically link the volume visualization to the navigation of the cell lineage tree, the application took advantage of volume rendering multi-component datasets. A two-component dataset was produced by using as first components the intensity values of the original microscopy image and as second components the cell identifiers produced by AceTree for that specific stage of embryo development.

In order to produce the second component, individual cells were segmented from the confocal microscopy images and then labeled using the cell identifiers provided by the results of the cell lineage tracing of AceTree.[3,4] Since the cell identifier labels are co-registered to the original microscopy images it is then possible to modulate the transfer functions of

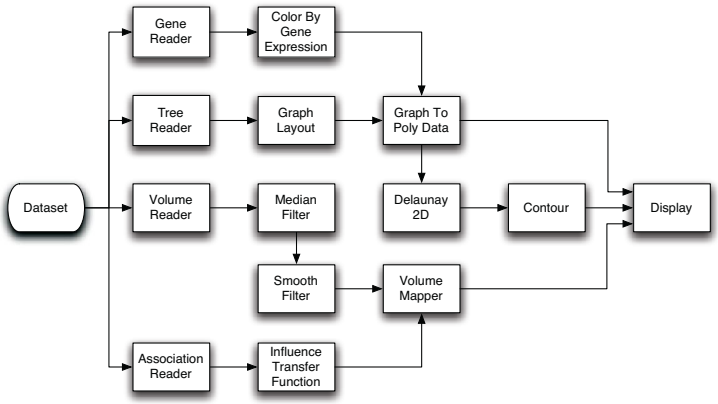

Figure 2. Visualization pipeline of the application.

the volume rendering based on the values of the cell identifiers. Through this mechanism, specific cells can be highlighted or removed from the volume visualization.

Although AceTree already performs the segmentation of the datasets and reconstructs the cell lineage, the intermediate images resulting from the cell labeling process were not available as output of the processing. Therefore, we reproduced the cell segmentations by using image analysis components from the Insight Toolkit (ITK). The image analysis pipeline used is described in Figure 4.

First the confocal microscopy datasets where smoothed using five iterations of a median filter that took into account the strong anisotropy of the pixel dimensions, a factor of almost $1:10$. The resulting smoothed images have most of the noise removed and present a consolidated region of almost uniform intensity inside the cells. These smoothed images were then thresholded to be used as the speed image for a Fast Marching level set filter.[7,21] The fast marching filter was run once for each one of the cells identified by AceTree for the current stage of embryo development. At each one of these iterations, the cell coordinates were used as seed point for the fast marching filter. The output of the fast marching filter was then labeled using the cell identifier produced by AceTree. A single image with the consolidation of all the labels corresponding to the cells of the current developmental stage was generated. The original microscopy image was merged with the image of labels in order to produce a two-components datasets. By creating independent transfer functions for each cells, it was possible to highlight individual cells according to selections made in the cell lineage tree. The volume mapper performs a ray-casting through the volume of data and for each sample point applies the appropriate transfer function. An example of individual cells hightlighting is presented in Figure 5.

4.1.3 Interaction

The application provides several interaction scenarios that facilitate the exploration of cell lineage, spatial correspondence of the cells, and gene expression. The simplest and most obvious interaction is zoom and pan in the tree layout and zoom, pan, and 3D rotation in the volume display. When zooming in the tree layout, the system exposes more cell annotations. Similarly, when zooming out the tree layout, the cell annotations will hide to improve visibility.

In the tree widget, the user can expand the subtrees to expose the cells in the subtrees. In addition, similar collapsing of subtrees can be performed in the tree layout. By selecting the cell, the subtree rooted in that cell will be hidden, while the cell will still be visible. The cell with the hidden subtree will be marked as such. An example of collapsing subtree in the tree layout view can be seen in the Figure 3 (g).

Finally, there is an elaborate selection system that links all views. A slider is provided to the user to select the time of embryonic development. This will load appropriate volume data corresponding to that time and draw a line on the tree layout at that stage of development. This can be seen in (e) and (f) of Figure 3, and in Figure 6, where the highlighted lines show the current development time. Figure 3 (i) shows the whole application with the selected volume time loaded.

Another selection mode allows user to select the individual cells, groups of cells, or subtrees of cells. Once the user selects a set of cells in the tree widget, or the tree layout, the same set is selected in the other tree view, as well as highlighted

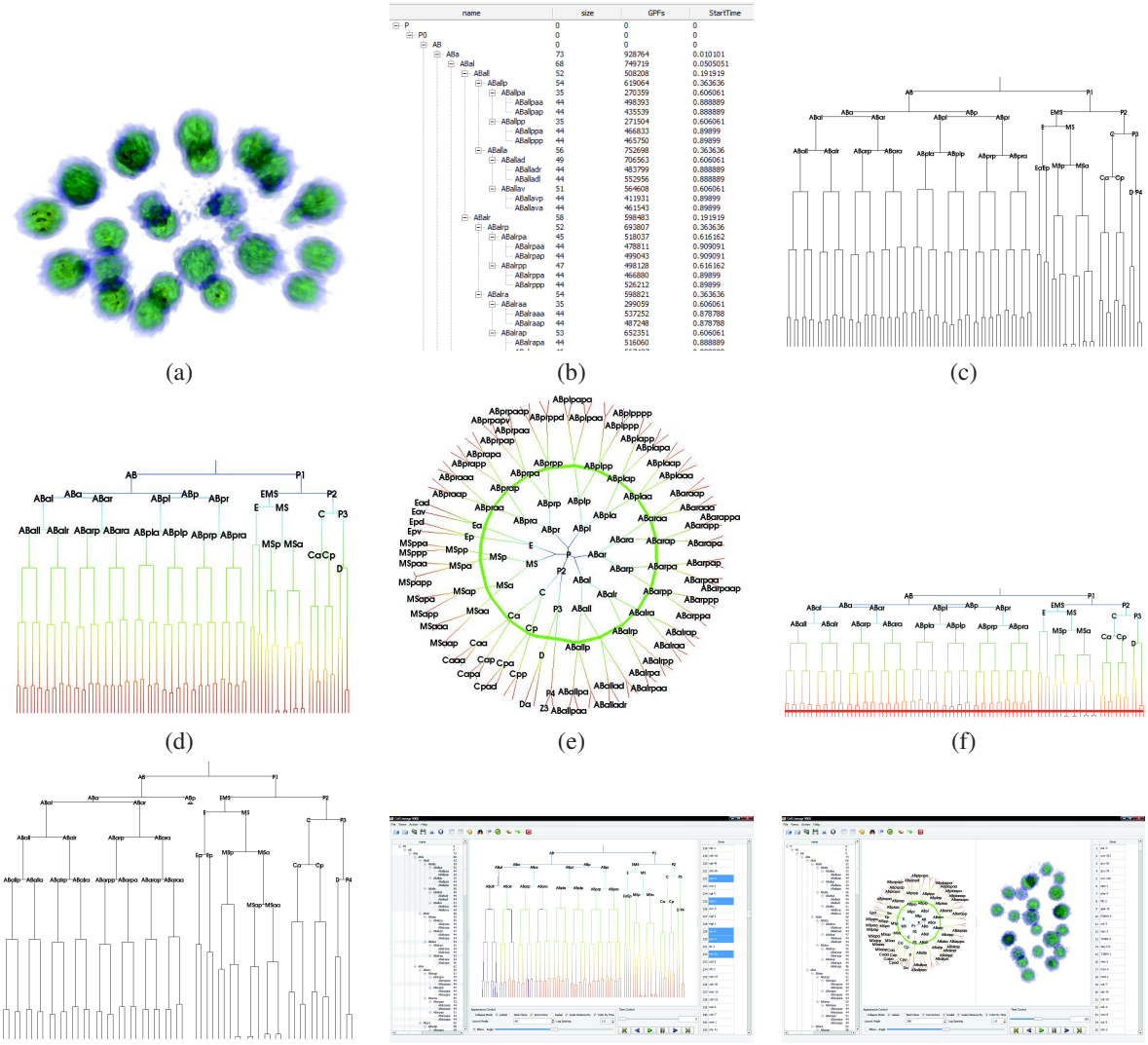

Figure 3. (a) Volume view. (b) Standard tree view available for quick access to individual cells. (c) Vertical tree layout view with elbow links. (d) Vertical tree layout view, colored by the development time. (e) Radial layout tree view with straight links. (f) Vertical layout tree view with elbow links colored by the development time and the time line for specific development stage. (g) Tree Layout view with the collapsed subtree rooted in **ABp**. (h) Linked selection. (i) Full program.

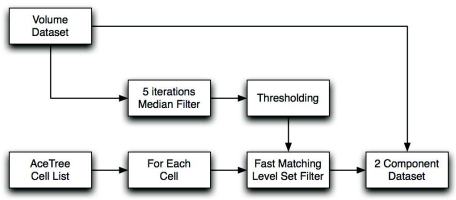

Figure 4. Image analysis pipeline used for generating dual-components datasets suitable for highlighting cells during volume visualization.

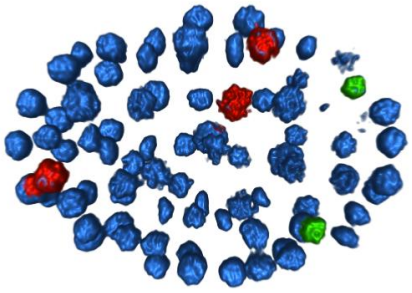

Figure 5. Volume visualization of the embryo developmental stage at 90 minutes, with cells **Eav**,**MSapp** and **ABaraapa** highlighted in red, and cells **Capp** and **ABprpapd** highlighted in green. This diagram illustrates how the volume visualization can be linked to selections made on the cell lineage tree display.

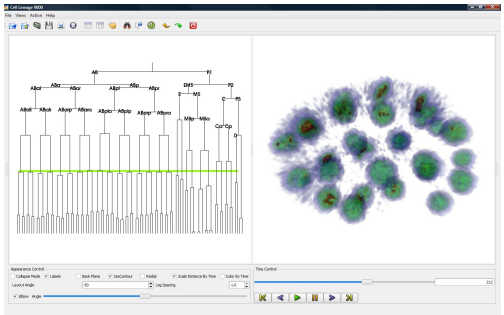

Figure 6. Tree Layout view and Volume view with the same time selected.

in the volume display. In addition, the genes in the gene expression display are highlighted according to the set of selected cells. The user can also select the gene set in the gene expression display and cells corresponding to these genes will be selected in other views. Figure 3 (h) shows example where several cells and subtrees of cell lineage are selected in the tree widget and corresponding cells are selected in the tree layout.

4.2 Scalability

The initial prototype was developed using a serial pipeline. This means the largest datasets the application can handle is in the order of thousands of cells and the image size in the order of 10 million voxels. For larger datasets, the combination of data parallelism and parallel rendering has to be employed. VTK natively supports data streaming. That said with the technologies in ParaView, the application can render large volumes in parallel.

The approach taken for the large data is to distribute data spatially and based on the cell lineage subtree the cells appear in. Currently ParaView can already process and render the cell lineage in parallel except of the layout of the actual tree. There are several ways to layout the tree in parallel but for the purpose of this application the actual parallel layout was not developed. It will be however, explored in future work.

4.3 Reproducibility

Reproducibility is a fundamental aspect of scientific work. More fundamental even than the peer-review process. After all, peer-review is just a method for verifying the reproducibility of published technical work. Unfortunately, the practical possibility of reproducing every experiment that is reported in the literature is hampered by the amount of technical resources that are needed for running such experiment. In the fields of image analysis and visualization, the software products themselves account for the most expensive resource required to reproduce technical work. The second most costly resource is the access to the original data used by the authors.

The current availability of open source toolkits, such as VTK and ITK, reduces the amount of effort that readers and reviewers must invest in order to reproduce published work. Further improvements can be made when authors share their

source code and data along with the text of their papers. In the preparation of this paper we benefited largely from the fully reproducible publications that distributed not only the original microscopy data of *C. elegans* embryos but also the executables of the AceTree and StarryNight applications used for generating cell lineages.[3,4,8] Thanks to the openness of these authors, we were able to reproduce their results with minimal effort, and then we were able to focus on applying the novel resources of the VTK information visualization framework.

In order to keep up to the standards that made this paper possible, we are also providing the source code used to produce results presented in this paper. This material can be downloaded from http://public.kitware.com/pub/vtk/doc/.

5. FUTURE WORK

There are several aspects of this application that should be addressed in the future. They include improved handling of even larger datasets, adding the capability of comparing lineage trees, and adding an option to search for motives in the branching pattern of lineage trees.

Currently the volume rendering is parallelized, but the tree algorithms are not. When exploring higher organisms, such as zebrafish, the number of cells can be a few orders of magnitude higher. The *C. elegans* lineage has around one thousand cells while analysis of the embryonic development of zebrafish will require analysis of up to a million cells. These much larger trees would benefit from parallelized rendering and traversal.

With higher organisms, the cell lineage is typically stereotypic but not completely deterministic, so the option of comparing trees would be extremely useful. This ability will be essential to define exactly how the lineage normally varies from organism to organism and what patterns are conserved. Furthermore, searching for motives in the branching would provide researchers with a powerful tool to detect mutations and find patterns. Figure 7 shows example mutations from *C. elegans* which result in sterotypic alterations in the pattern of the lineage tree. The normal lineage tree is shown on the left. Certain mutations will cause a level of the lineage tree to be skipped while other mutations will cause a level of the tree to be repeated over and over. As can be seen in the figure the nature of these motifs can be fairly complex. Lineage motifs also exists in higher organisms in the context of stem cells. Some stem cells follow a motif where one daughter cell remains a stem cell while the other daughter cell differentiates either immediately or after a few rounds of cell division which serve to expand the number of differentiated cells. Using tree comparison or motif searching will allow normal patterns in lineage trees to be identified as well as allow changes in these patterns caused by genetic mutations to be detected.

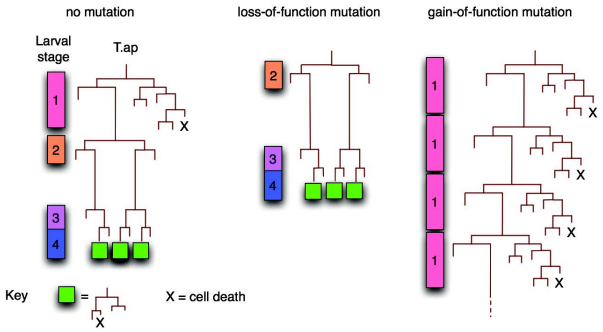

Figure 7. Lineage Mutation

6. CONCLUSION

This paper describes an application for analysing the embryogenesis using the combination of information and traditional visualization. The application uses the new Titan informatics toolkit of the Visualization toolkit to display cell lineage trees, and volume visualization of VTK to display the confocal microscopy images. By linking the selection from the tree views, volume view, and the gene expression view, it provides the research several tools required for effective analysis. The algorithms used to perform the visualization are scalable to much larger models, so it is feasible to envision analysing much more complex embryos. Most examples in this paper were performed on the dataset of *C. Elegans* embryos. However, the application will be extended to perform the same operations on a much larger datasets of zebra fish. In addition, because

the application is based on open source, general purpose toolkits, it can be readily extended to perform other analyses, such as searching for motives, gene expression visualization, and cell lineage comparison.

ACKNOWLEDGMENTS

This work was done in part at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. The authors would also like to thank Lisa Sobierajski Avila for her assistance with volume rendering aspects of the project.

REFERENCES

1. J. Sulston and H. Horvitz, "Post-embryonic cell lineages of the nematode, caenorhabditis elegans," *Dev Biol* **56**, pp. 110–56, March 1977.
2. J. Sulston, E. Schierenberg, J. White, and J. Thomson, "The embryonic cell lineage of the nematode caenorhabditis elegans," *Dev Biol* **100**, pp. 64–119, Nov 1983.
3. Z. Bao, J. Murray, T. Boyle, S. Ooi, M. Sandel, and R. Waterston, "Automated cell lineage tracing in caenorhabditis elegans," *Proc Natl Acad Sci U S A* **103**, pp. 2707–12, Feb 2006.
4. T. Boyle, Z. Bao, J. Murray, C. Araya, and R. Waterston, "Acetree: a tool for visual analysis of caenorhabditis elegans embryogenesis," *BMC Bioinformatics* **7**, Jun. 2006.
5. S. Megason and S. Fraser, "Digitizing life at the level of the cell: high-performance laser-scanning microscopy and image analysis for in toto imaging of development," *Mech Dev* **120**, pp. 1407–20, Nov 2003.
6. S. Megason, A. Amsterdam, N. Hopkins, and S. Lin, *Green Fluorescent Protein: Properties, Applications and Protocols, Second Edition. Edited by M Chalfie and SR Kain*, ch. Uses of GFP in Transgenic Vertebrates. Wiley Press, 2006.
7. L. Ibanez, W. Schroeder, L. Ng, and J. Cates, *The ITK Software Guide*, Kitware, Inc., 2003.
8. L. D. Stein, Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D'Eustachio, D. H. A. Fitch, L. A. Fulton, R. E. Fulton, S. Griffiths-Jones, T. W. Harris, L. W. Hillier, R. Kamath, P. E. Kuwabara, E. R. Mardis, M. A. Marra, T. L. Miner, P. Minx, J. C. Mullikin, R. W. Plumb, J. Rogers, J. E. Schein, M. Sohrmann, J. Spieth, J. E. Stajich, C. Wei, D. Willey, R. K. Wilson, R. Durbin, and R. H. Waterston, "The genome sequence of caenorhabditis briggsae: A platform for comparative genomics," *PLoS Biology* **1**, Nov. 2003.
9. I. Letunic and P. Bork, "Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation," *Bioinformatics* **23**, pp. 127–128, 2007.
10. C. North, "Multiple views and tight coupling in visualization: A language, taxonomy, and system," in *CSREA CISST Workshop of Fundamental Issues in Visualization*, pp. 626–632, June 2001.
11. T. Pattison and M. Phillips, "View coordination architecture for information visualization," in *2001 Asia-Pacific symposium on Information visualization*, pp. 165–169, Australian Computer Society, Inc., (Darlinghurst, Australia), 2001.
12. "Prefuse information visualization toolkit." http://www.prefuse.org.
13. "GraphViz graph visualization software," http://www.graphviz.org.
14. "The InfoVis toolkit," http://ivtk.sourceforge.net.
15. B. Wylie and J. Baumes, "The Titan informatics toolkit." Not yet published.
16. Kitware, *The VTK User's Guide*, Kitware, Inc., 2003.
17. *VTK: The Visualization Toolkit*, 2003.
18. A. Cedilnik, B. Geveci, K. M. J. Ahrens, and J. Favre, "Remote Large Data Visualization in the ParaView Framework," in Raffin *et al.*,[20] pp. 163–170.
19. A. Squillacote, *ParaView Guide, A Parallel Visualization Application*, Kitware Inc., 2005.
20. B. Raffin, A. Heirich, and L. P. Santos, eds., *Eurographics Symposium on Parallel Graphics and Visualization*, (Braga, Portugal), Eurographics Association, 2006.
21. J. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge University Press, 1996.